MÉTODOS E TÉCNICAS PARA USAR A INTERNET DIRETAMENTE COMO CORPUS: O CASO DOS DICIONÁRIOS ON-LINE DE ESPANHOL VALLADOLID-UVa[1,2]

Sven Tarp[3]
Pedro A. Fuertes-Oliveira[4]
Tradução: Luísa Rabaldo[5]
Revisão: Ana Eliza Pereira Bocorny[6]; Sandra Loguercio[7]

Resumo: Inicialmente, este artigo aborda algumas das consequências que o desenvolvimento tecnológico trouxe para a lexicografia, principalmente quanto aos diferentes tipos de bases empíricas que podem ser usados em projetos de dicionários. Em seguida, as principais vantagens e desvantagens de usar a Internet como corpus são listadas e comparadas com os resultados obtidos com corpora "tradicionais". Para ilustrar, o artigo mostra de que forma a Internet é usada como a principal fonte empírica para selecionar lemas e itens de significado para os Dicionários On-line de Espanhol Valladolid-UVa. Os métodos e ferramentas empregados no projeto são discutidos juntamente com as competências, conhecimentos e habilidades dos lexicógrafos. Por fim, o artigo fornece algumas conclusões gerais, bem como faz recomendações e levanta hipóteses para futuros trabalhos e pesquisas na área da lexicografia.

Palavras-chave: lexicografia da internet; lexicografia *online*; metodologia lexicográfica; base empírica; seleção do lema; seleção do significado; bases de dados lexicográficos; dicionários de espanhol; dicionários de língua geral.

1. Introdução

Por meio da análise do processo geral de compilação de dicionários descrito por Fuertes-Olivera e Tarp (2014: 85), conclui-se que há três casos em que os lexicógrafos

[1] N.A.: Este artigo é uma adaptação de um texto publicado em *Lexikos 26* (TARP; FUERTES-OLIVEIRA, 2016). N.E.: A presente tradução foi autorizada para ser publicada em português pelo autor Sven Tarp.

[2] Tradução revisada pelo autor Sven Tarp e título modificado conforme solicitação do autor.

[3] Centro de Lexicografia, Universidade de Aarhus, Dinamarca e Departamento de Africâner e Dinamarquês, Universidade de Stellenbosch, África do Sul (st@cc.au.dk).

[4] Centro Internacional de Lexicografia, Universidade de Valladolid, Espanha e Departamento de Africâner e Dinamarquês, Universidade de Stellenbosch, África do Sul (pedro@emp.uva.es).

[5] Bacharelanda em Letras Português/Inglês e bolsista de Iniciação Científica, Instituto de Letras, UFRGS.

[6] Professora do Departamento de Línguas Modernas e do Programa de Pós-Graduação em Letras, Instituto de Letras, UFRGS.

[7] Professora do Departamento de Línguas Modernas e do Programa de Pós-Graduação em Letras, Instituto de Letras, UFRGS.

podem necessitar de acesso a dados empíricos para fazer um bom trabalho. O primeiro é quando procuram informações sobre as necessidades lexicográficas do público-alvo, com o objetivo de definir uma concepção de dicionário que possa responder às suas necessidades; o segundo é quando selecionam e preparam os dados lexicográficos para inclusão em um dicionário; e o terceiro é quando avaliam a utilidade do dicionário em termos de satisfação do usuário. Na verdade, há também uma quarta situação em que dados empíricos externos podem ser necessários: quando lexicógrafos avaliam o mercado para determinar as possibilidades de vendas do produto. O quarto caso, entretanto, está mais relacionado ao lado comercial do projeto do que aos aspectos lexicográficos, em um sentido estrito. De qualquer forma, para cada uma dessas situações existe um conjunto de métodos mais ou menos apropriados, isto é, mais ou menos confiáveis e rápidos em termos de produtividade e qualidade do produto final.

Em seguida, veremos as bases empíricas e os métodos correspondentes que podem ser aplicados ao selecionar lemas e itens de significado (sentido) em um projeto lexicográfico *on-line*. Discutiremos algumas das vantagens e desvantagens mais importantes ao usar a Internet como um *corpus* e compararemos com os resultados obtidos com *corpora* "tradicionais" de texto. Como exemplo, utilizaremos um projeto que está sendo realizado atualmente no Centro Internacional de Lexicografía da Universidade de Valladolid, chamado *Diccionarios en Línea de Español "Universidad de Valladolid"*, referidos neste artigo como Dicionários On-line de Espanhol Valladolid-UVa. O projeto, que originalmente foi iniciado como uma colaboração entre o Centro de Valladolid e o Centro de Lexicografia da Universidade de Aarhus, baseia-se na teoria funcional da lexicografia e é inspirado em um projeto dinamarquês similar (ver Fuertes-Olivera; Bergenholtz, 2015). Por fim, apresentaremos os resultados obtidos até o presente momento e abordaremos a necessidade de contar com a intuição como um método intangível, mas muito relevante e inevitável no desenvolvimento de dicionários.

2. A relação entre lexicografia e tecnologia

Em uma perspectiva histórica, pode-se observar uma relação íntima e complexa entre a lexicografia e a tecnologia. Isso implica, entre outras coisas, que o desenvolvimento tecnológico pode não apenas criar novas ferramentas para auxiliar lexicógrafos na realização de tarefas, mas também novas bases empíricas a partir das quais podem recuperar dados, bem como depararem-se com a necessidade e a possibilidade de desenvolverem novos métodos de trabalho. A reflexão é particularmente relevante em períodos históricos como o atual, em que novas tecnologias disruptivas estão sendo introduzidas na lexicografia com consequências que ainda não são totalmente compreendidas:

> Atualmente estamos no meio de uma nova transição da base material e tecnológica da lexicografia com a introdução de novas ferramentas e métodos de produção, bem como novas plataformas

> e meios para apresentar o produto lexicográfico e o uso extensivo de *corpora* para a coleta de material. O desenvolvimento e a inovação tecnológica estão mais acelerados do que nunca. (...) Sabemos o ponto de partida, mas temos apenas uma vaga ideia de onde chegaremos. (Gouws; Tarp, 2017)

Em geral, há uma variedade de fontes das quais os lexicógrafos podem obter dados. Bergenholtz e Tarp (1995: 90-96) discutem, dentre as mais importantes, a introspecção, a multispecção, os especialistas externos, os dicionários existentes, os manuais, os livros didáticos, os cartões de exemplo e os *corpora* textuais. Com exceção das três primeiras, essas fontes empíricas só são possíveis graças ao desenvolvimento tecnológico em vários estágios: invenção do papel, canetas, encadernação de livros, máquinas de impressão, computadores e bancos de dados. Desde então, com a introdução e o desenvolvimento da tecnologia *on-line*, outra fonte de dados empíricos foi colocada à disposição da lexicografia: a Internet.

É interessante notar que Bergenholtz e Tarp (1995) discutem a consulta a especialistas externos como uma forma de multispecção e o uso do próprio conhecimento como uma forma de introspecção. Pode ser que tenham razão, mas parece que há uma diferença entre o uso da introspecção em termos de habilidades e competências linguísticas, como é normalmente compreendido dentro da linguística, e o uso de conhecimento especializado e atualizado que está armazenado na memória de alguém. Como Tarp (2008: 131-136) argumentou, no desenvolvimento de vários tipos de dicionários, é importante distinguir, de um lado, as habilidades linguísticas e, de outro, o conhecimento erudito de uma língua, tal como fornece, por exemplo, a teoria linguística. A esse respeito, parece lógico também distinguir o uso da competência de linguagem e o uso de conhecimento especializado no processo de compilação de dicionários. Assim, embora exista certa confusão terminológica na literatura lexicográfica existente, a introspecção - em vez de ser uma base empírica em si - deve ser considerada um método para a obtenção de tipos específicos de dados empíricos. Trata-se de um método feito para "olhar" para si mesmo, a fim de adquirir material para diferentes propósitos. As bases empíricas "internas" das quais os lexicógrafos podem se valer por meio desse método são as competências, habilidades e conhecimentos de linguagem, aos quais pode-se acrescentar a experiência pessoal em geral.

As diversas fontes empíricas raramente são usadas sozinhas. Em uma resenha de livro, Kilgarriff (2012) fornece um exemplo sobre como dois tipos diferentes de bases empíricas combinam-se:

> Eu notei um erro lexicográfico. Nas páginas 211-213, temos uma análise do verbo frasal em inglês *call back*. Seis significados são apresentados e, no sexto, surge o exemplo "*I cannot call his face back.*" Como falante nativo de inglês, esse exemplo me causou

> estranheza. Essa frase está completamente errada. (Poderíamos dizer *I cannot recall his face*). Uma breve pesquisa revelou que a tal "frase exemplo" existe em vários dicionários e ferramentas de tradução: um erro dicionarizado que foi copiado e recopiado de dicionário para dicionário. (Kilgarriff 2012: 28)

Ao dizer que "esse exemplo me causa estranheza" como falante nativo de inglês, Kilgarriff sugere que sua competência na língua materna o alerta sobre um possível problema, confirmado e explicado por meio da consulta de outras bases empíricas, neste caso, dicionários existentes e ferramentas de tradução. Esse é claramente o método correto a ser aplicado em tais situações, uma vez que "a fonte primária de evidências do lexicógrafo sobre como uma palavra é usada muda de subjetiva para objetiva; da introspecção à observação dos contextos" (Kilgarriff 1997: 111). Nota-se que Kilgarriff fala aqui da fonte primária de evidência, que não é a única, embora erre ao definir introspecção e observação dos contextos como fontes, na medida em que ambos são métodos para acessar as fontes reais de evidência.

3. Corpus *vs.* Internet: discussão preliminar

Os primeiros *corpora* compostos com textos eletrônicos foram introduzidos na década de 1960, e não pararam de crescer desde então. As duas primeiras décadas após o surgimento foram caracterizadas por uma significativa batalha de ideologias entre os pesquisadores que defenderam a relevância dos *corpora* tanto para a linguística como para a lexicografia, e aqueles que se opuseram a essa visão com vários argumentos, geralmente defendendo a introspecção como método mais apropriado para a obtenção de material empírico. Um dos defensores da introspecção foi Lees (1962) que declarou:

> Você é um falante nativo de inglês; em dez minutos é capaz de produzir mais ilustrações de qualquer questão gramatical da língua inglesa do que achará em milhões de palavras em textos aleatórios. (Lees 1962: 110)

Aos poucos, a discussão desapareceu. Meio século após sua introdução, não resta mais nenhuma dúvida de que *corpora* de textos eletrônicos podem ter grande valor não apenas para a pesquisa linguística, mas também para os lexicógrafos ao realizarem uma série de tarefas relacionadas à compilação de dicionários. Isso tem sido debatido por vários estudiosos envolvidos na lexicografia prática, entre eles Bergenholtz (1996), Atkins e Rundell (2008) e Hanks (2012). A prova do pudim é a existência de muitos dicionários de alta qualidade que foram compilados fazendo uso desse tipo de base empírica (ver, por exemplo, Sinclair, 1997), embora a ânsia tenha ido, por vezes, longe demais, pelo menos quanto à seleção de termos e definições em dicionários especializados (ver Tarp, 2016, e Xue; Tarp,2016).

No entanto, como consequência negativa desse desenvolvimento positivo, a introspecção como método de uso das competências e conhecimentos próprias de um indivíduo é, por vezes, subestimada ou até mesmo ignorada. Embora "lexicógrafos nunca devam confiar somente na abordagem introspectiva" (Bergenholtz; Tarp 1995: 92), especialmente em casos dúbios, frequentemente esquece-se que a introspecção sempre existe como um filtro na base das escolhas do lexicógrafo, na medida em que nenhum dicionarista introduziria dados linguísticos ou de qualquer outro tipo com os quais discorda – como a estranheza causada em Kilgarriff – sem antes negociar sua correção com outras fontes empíricas.

Hoje em dia, *corpora* compostos de textos contendo centenas de milhões de palavras estão disponíveis aos compiladores de dicionários. O *Big Data*, por exemplo, já é uma realidade, mas o entusiasmo gerado por esse desenvolvimento não deve ofuscar o fato de que nenhum *corpus*, por maior que seja, pode competir com a enorme coleção de textos e palavras que podem ser acessados através da Internet. O desenvolvimento de métodos que permitam o uso dessa base empírica quase ilimitada constitui, sem dúvida, um desafio cada vez mais relevante para a lexicografia.

Segundo Fuertes-Olivera (2014: 51), um *corpus* lexicográfico, ou seja, um *corpus* que pode ser usado para auxiliar na criação de dicionários, pode ser definido como "qualquer coleção de textos em que os lexicógrafos possam encontrar inspiração para completar as estruturas das quais precisam na construção de um dicionário real". Como já mencionado, a Internet é composta por uma coleção de textos. Assim, se um lexicógrafo pode encontrar inspiração nessa grande coleção de textos, a Internet também pode ser considerada um tipo de *corpus* lexicográfico de acordo com a definição acima. Esse é também o ponto de vista de Kilgarriff e Grefenstette (2003: 334) que dizem que "a resposta à questão 'a Internet é um corpus?' é sim."

Assim, há duas maneiras diferentes de usar a Internet em um projeto lexicográfico, a saber: 1) construir um *corpus* de textos encontrados na Internet, e 2) usar a Internet diretamente como um *corpus* - em ambos os casos por meio de ferramentas de busca e outros recursos. Cada um desses dois tipos de *corpus* lexicográfico tem suas vantagens e desvantagens. Abaixo listamos algumas das vantagens do uso da Internet como *corpus*, em relação ao uso de *corpora* "tradicional" composto de coleções de textos, sejam coletados da Internet ou de outro lugar:

- Os lexicógrafos têm acesso a um número muito maior de textos do que os incluídos em qualquer *corpus* de textos selecionados.
- Os textos estão sempre atualizados.
- Tempo e dinheiro são poupados quando não é necessário compor um *corpus* (que é um requisito em relação a tipos específicos de dicionários, principalmente os especializados).
- O processo de busca pode ser facilmente limitado a áreas geográficas específicas, um fato importante para um idioma multinacional como o espanhol.

- O uso da Internet pode levar à identificação e à seleção de mais itens de significado do que aquelas que podem ser encontradas em um *corpus* fabricado.

Quanto às desvantagens de usar a Internet como *corpus*, os seguintes aspectos parecem ser os mais importantes:

- Nem a qualidade nem a origem dos textos podem ser controladas.
- Em alguns textos, os autores podem não ser pessoas reais.
- Os autores podem ter um baixo nível de proficiência na língua em questão.
- Os textos podem não ter sido revisados e corrigidos.
- É difícil calcular a frequência dos fenômenos linguísticos que aparecem nos textos.

Algumas das desvantagens citadas acima podem não ser relevantes para projetos concretos de dicionários. Gudmann (2014: 32), por exemplo, argumenta que "as informações sobre frequência (...) não são particularmente relevantes para um dicionário geral monolíngue". Em outros casos, as desvantagens podem ser neutralizadas ou, ao menos, consideravelmente reduzidas, por um lexicógrafo bem preparado que desempenhe um papel ativo com base em sua competência linguística, suas habilidades, seu conhecimento e sua experiência. Retomaremos essa questão mais adiante. Até aqui, nossa conclusão preliminar é que, apesar das inegáveis desvantagens, é perfeitamente possível, e até mesmo benéfico, usar a Internet como principal fonte empírica, sem recorrer aos *corpora* "tradicionais" de textos, quando o objetivo é a produção de dicionários com qualidade ainda maior.

4. Selecionando lemas

Nós, sinceramente, duvidamos que o *corpus* tradicional, composto por uma coleção de textos, seja a fonte empírica mais apropriada para a seleção de lemas, principalmente se essa seleção for feita desde o início. Até onde sabemos, os grandes dicionários de língua geral que usam *corpora* para esse propósito são, em sua maioria, dicionários que tinham um estoque de lemas selecionado antes da introdução dos *corpora,* que agora são usados "apenas" para fornecer lemas adicionais, entre outros dados. Um método diferente e uma base empírica são, portanto, necessários quando o desafio é uma seleção rápida e confiável de lemas para um projeto lexicográfico completamente novo da magnitude dos Dicionários On-line de Espanhol Valladolid-UVa, planejados para lidar com mais de cem mil lemas e um número muito maior de significados. A principal base empírica escolhida para esse projeto foi, portanto, a Internet, acompanhada de um método que será descrito nesta seção.

A ideia básica é que a Internet já contém um número considerável de listas de palavras menores ou maiores, com acesso e uso gratuitos. O desafio é, portanto, encontrar essas listas e fazer uso delas. Isso é feito por meio de um rastreador de Internet projetado especificamente para esse fim pela empresa dinamarquesa Ordbogen.com, que, devido ao seu modelo de negócios, é o provedor de dicionários *on-line* em assinatura mais bem-sucedido do mundo atualmente.

Assim que determinadas listas de palavras úteis são encontradas pelo rastreador de Internet, essas listas são copiadas e coladas em um chamado carregador de lema (veja a Figura 1), outra ferramenta desenvolvida pela Ordbogen.com e concebida pelo professor Henning Bergenholtz do Centro de Lexicografia de Aarhus. O carregador de lema atribui automaticamente um lema a um cartão no banco de dados e tem a vantagem de não duplicar a entrada, rejeitando-os caso já constem no banco de dados.

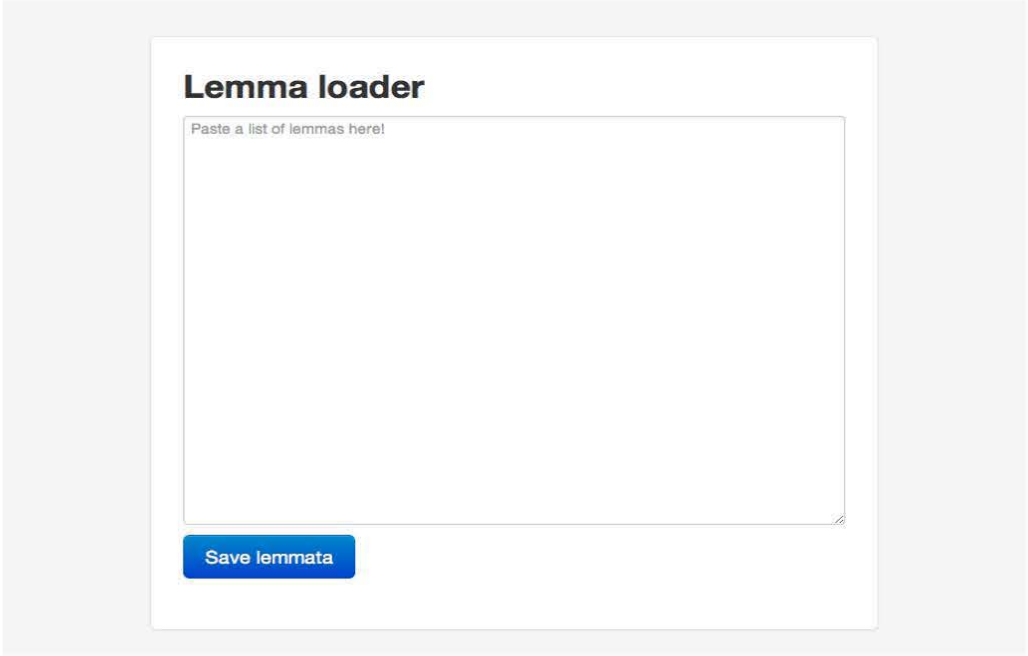


Figura 1: Captura de tela do carregador de lema com o campo em que as listas de palavras copiadas são coladas.

As experiências mostram que o método para selecionar lemas por meio de um rastreador de Internet e um carregador de lema é muito eficiente, rápido e totalmente confiável no caso de lemas compostos por uma palavra. Isso é revelado pelo fato de que, apenas um mês após o início da seleção de lemas para os Dicionários On-line de Espanhol Valladolid-Uva (julho de 2013), o banco de dados já continha 58.000 cartões com lemas de uma palavra.

O passo seguinte no processo de seleção de lema é a revisão manual que ocorre quando os dados gramaticais formais são anexados ao cartão do lema. A revisão é realizada pelo editor-chefe, Pedro A. Fuertes-Olivera. Devido às características dos dicionários e à capacidade de armazenamento quase ilimitada do banco de dados, o projeto não trabalha com critérios de inclusão de lema, apenas com critérios de exclusão. Com essa abordagem, apenas os lemas que apresentam claramente erros ortográficos ou que não podem ser documentados na base empírica, isto é, na Internet, são excluídos, mesmo quando são encontrados em algumas listas de palavras antigas, mas não na Internet como tal. Até agora foram poucos os casos em que um lema teve de ser excluído, o que também aponta para a eficiência do método.

A fim de garantir um tratamento sistemático da língua, após a seleção inicial foram elaboradas várias listas temáticas com cores, números, cidades de determinado tamanho, rios de mais de 1.000 km, etc., e as palavras correspondentes introduzidas no banco de dados como lemas. Esse trabalho durou de setembro a novembro de 2013 e resultou na inclusão de mais 10.000 lemas.

Além das fontes empíricas mencionadas, outras fontes também são usadas para fornecer um fluxo de novos lemas. Por exemplo, quando os lexicógrafos estão consultando a Internet para identificar itens de significado para os lemas selecionados (ver Seção 5), eles detectam simultaneamente um número considerável de sinônimos, antônimos e combinações de palavras que são continuamente introduzidos no banco de dados como novos lemas.

Por fim, há a questão de expressões idiomáticas e outras expressões fixas que são geralmente selecionadas como lemas independentes nos Dicionários On-line de Espanhol Valladolid-UVa. Neste caso, há quatro fontes: 1) No processo de detecção de itens de significado, ocasionalmente aparecem expressões fixas. Nesse caso, elas são enviadas ao editor-chefe, que as analisa e avalia pesquisando na Internet. 2) Dicionários existentes também são usados como fontes, 3) assim como o CREA Corpus composto e publicado pela *Real Academia Española*, de uso livre. 4) Por fim, várias expressões fixas também são encontradas em outras fontes, por exemplo, livros e artigos lidos pelos lexicógrafos relacionados a outras tarefas.

Como se pode ver, apenas as últimas três fontes de busca de expressões fixas não se apoiam na Internet como base empírica para a seleção de lemas para os Dicionários On-line de Espanhol Valladolid-UVa. Geralmente, o processo é muito rápido, eficiente e de baixo custo, o que, até agora, resultou em cerca de 20% mais lemas do que aqueles contidos nos dicionários espanhóis, até então, maiores.

5. Selecionando itens de significado

O método desenvolvido para selecionar itens de significado nos Dicionários On-line de Espanhol Valladolid-UVa é fortemente inspirado em um método similar usado no *Danish Internet Dictionaries* (ver Bergenholtz; Agerbo, 2014), mas tem algumas particularidades próprias. Grosso modo, o método de seleção de significado engloba as seguintes quinze etapas:

1. Um lema contido no banco de dados é escolhido na interface do usuário do lexicógrafo (ver Figura 2);
2. O botão "Google", localizado à esquerda na interface do lexicógrafo, é ativado;
3. Um resultado "tradicional" da pesquisa do Google é exibido (ver Figura 3);
4. As primeiras (3-20) páginas são ignoradas porque contêm apenas dados irrelevantes, em termos lexicográficos;
5. Os mini-textos exibidos em cada página são lidos para que se obtenha uma ideia geral do que se trata;

6. Usando o método "copiar e colar", as partes relevantes dos mini-textos são copiadas para um documento do Word;
7. Simultaneamente, colocações, exemplos, sinônimos, antônimos e combinações de palavras são selecionados para serem introduzidos nos respectivos campos no cartão representando o sentido em questão na interface do lexicógrafo (ver Figuras 4 e 5). Expressões idiomáticas e expressões fixas são enviadas ao editor-chefe para avaliação adicional;
8. Várias páginas do Google são revisadas até que não se encontre mais dados novos e tudo seja repetido. O número de páginas depende das características de cada lema, assim como da intuição do lexicógrafo, que se baseia na sua experiência;
9. Assim que uma quantidade satisfatória de dados empíricos é selecionada, os dados são agrupados de acordo com o significado;
10. Com base nos grupos de dados, as primeiras definições são escritas de acordo com as instruções lexicográficas preparadas pelo editor-chefe;
11. Nesta etapa, o lexicógrafo decide se está satisfeito, ou se é necessário repetir o processo inteiro ou parte dele para obter uma quantidade satisfatória de evidências empíricas;
12. Assim que o lexicógrafo termina a seleção de significados e redige as definições relativas a um lema, uma mensagem é enviada ao editor-chefe;
13. O editor-chefe revisa as definições e as compara com as que aparecem em quatro dicionários de espanhol previamente selecionados (ver Seção 6). Se alguma informação estiver faltando, pode ser que se inicie um novo processo de pesquisa, pois um princípio básico do projeto é que nenhuma definição seja copiada de outros dicionários;
14. Se as definições estiverem relacionadas a termos que ocorrem também na linguagem cotidiana, especialistas externos podem ser consultados para definir com mais exatidão;
15. Quando o editor-chefe estiver satisfeito - e outros dados relevantes, como gramática, sinônimos, antônimos, formações de palavras, colocações e exemplos forem incluídos - o lema é indicado para publicação *on-line*.

Até julho de 2016, o banco de dados dos Dicionários On-line de Espanhol Valladolid-UVa contém cerca de 40.000 cartões finalizados (cada um deles representando um significado), prontos para publicação. A experiência até agora mostra que, para 70% dos lemas, trabalhar com os mini-textos que aparecem como resultado da pesquisa do Google é suficiente. Para os 30% restantes, é necessário, ainda, acessar um ou mais *links* para encontrar dados adicionais em outros documentos. Neste último caso, os dados necessários são encontrados em 90% das situações. Apenas em 10% das ocorrências, o que representa 3% da totalidade dos lemas, é necessário realizar uma nova busca com uma variante do lema procurado. Isso significa que, para 97% de todos os lemas, uma única pesquisa do Google é suficiente para obter o material empírico necessário para a seleção de itens de significado e redigir definições de acordo com os critérios estabelecidos.

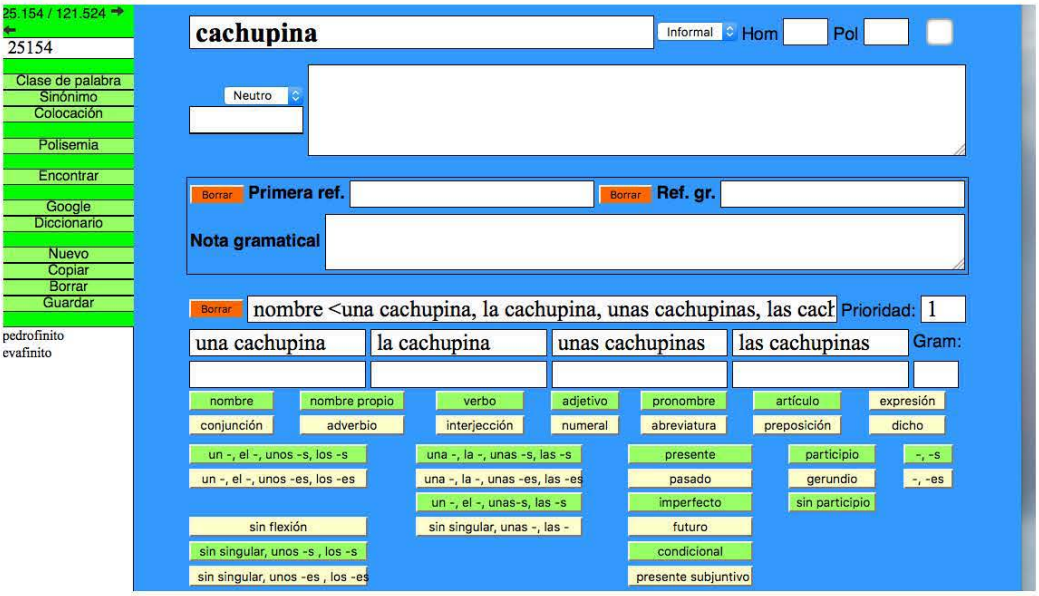


Figura 2: Interface do lexicógrafo da gramática da palavra "cachupina".

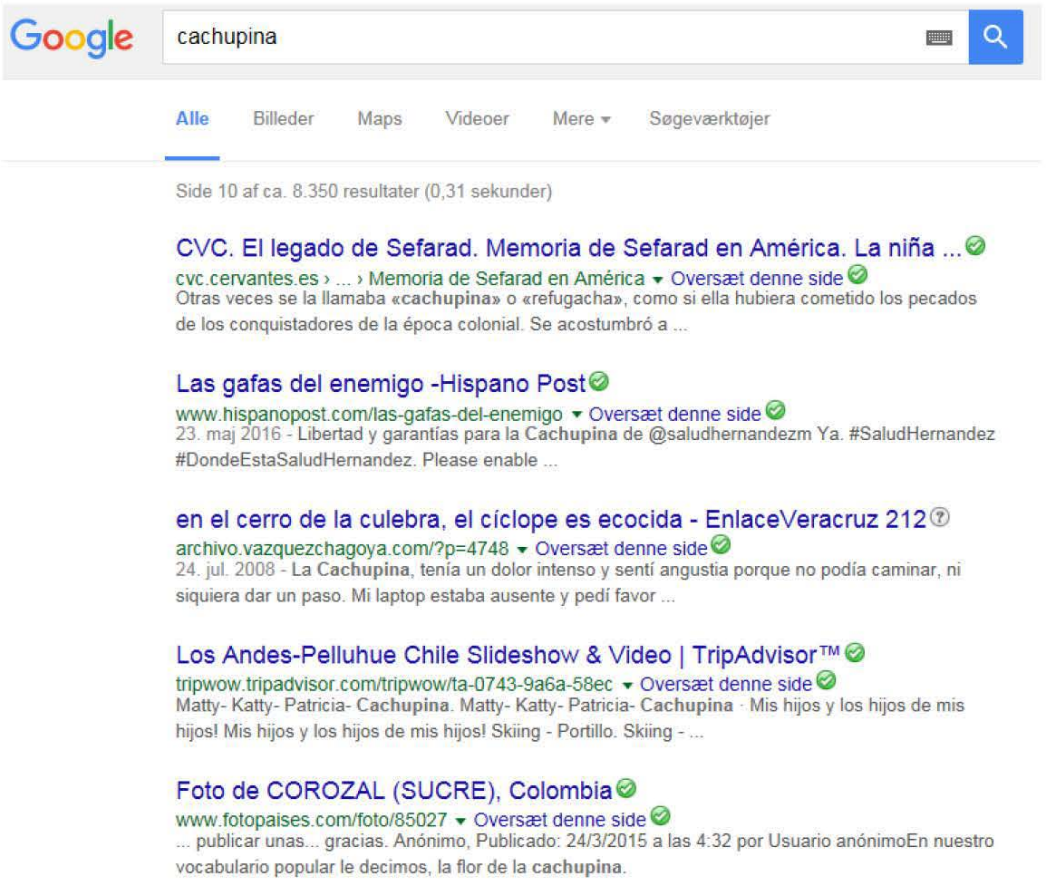


Figura 3: Resultado da pesquisa no Google pela palavra "cachupina".

Figura 4: Interface do lexicógrafo para a introdução de definição, sinônimos e antônimos para a palavra "cachupina".

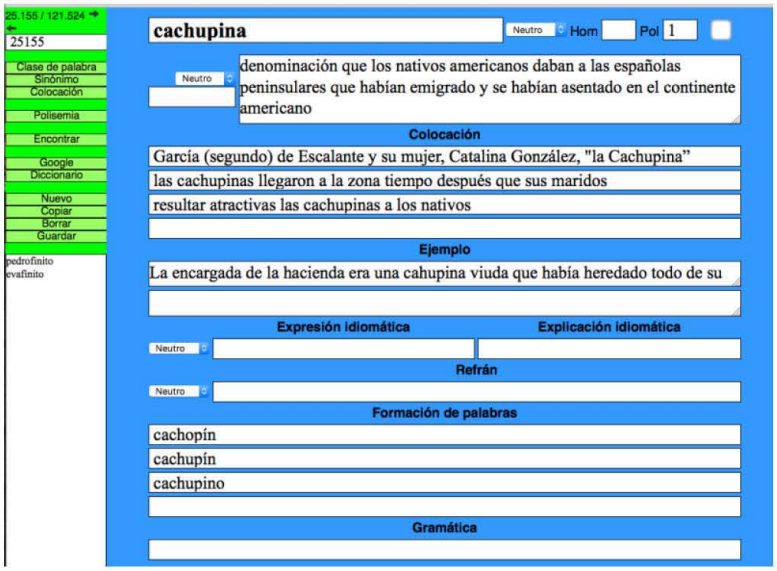


Figura 5: Interface do lexicógrafo para introdução de colocações, frases de exemplo, expressões fixas e formações de palavras para a palavra "cachupina".

6. Comparação com dicionários de espanhol semelhantes

Como mencionado acima, depois de redigir as definições dos diferentes significados, eles são comparados com aqueles encontrados em quatro dicionários de espanhol, a saber:

- María Moliner: Diccionario de Uso del Español (DUE)
- Aquilino Sánchez Pérez: Gran Diccionario de Uso del Español Actual (GDUEA)
- Manuel Seco: Diccionario del Español Actual (DEA)

- Real Academia Española: Diccionario de la Lengua Española (DLE)

Estes quatro dicionários estão entre os maiores e mais prestigiados dicionários de língua geral do espanhol. A comparação também serve como um tipo de controle de qualidade e indicação do que poderia ser melhorado (ver passo 13 na Seção 5). Até agora, a comparação tem sido favorável aos Dicionários On-line de Espanhol Valladolid-UVa, pois mostra que cada lema tratado com o método descrito tem uma média de 30-40% mais significados do que os lemas encontrados nos outros quatro dicionários.

A Tabela 1 mostra o número de significados que os cinco dicionários mencionados fornecem para oito lemas diferentes. Esses dados não são de forma alguma representativos, mas apenas um indicativo de como o método descrito na seção anterior, em alguns casos, pode gerar um número maior de significados.

Tabela 1: Comparação entre cinco dicionários de espanhol quanto ao número de significados de oito lemas selecionados.

Palavra	DUE	GDUEA	DEA	DLE	Valladolid
ababol	1	1	1	2	3
Cabila	1	1	2	2	3
Cable	4	4	4	6	11
cabestro	4	4	2	4	6
eclipsar	2	2	2	2	3
eclipsarse	2	2	2	3	4
halagar	4	2	2	4	3
machaca	5	5	5	6	15

A tendência refletida na Tabela 1 é corroborada por Gudmann (2015), que estudou cinco dicionários *on-line* de espanhol - entre eles a obra da Real Academia Española - e identificou um número surpreendentemente grande de lacunas de significado. Outra ilustração deste fenômeno é o tratamento das palavras *cachupín* e *cachupina*, que são usadas em algumas partes da América Latina e apresentadas como um único lema nos quatro dicionários mencionados acima, cada um deles com apenas um sentido. A Figura 7 mostra como eles são abordados pela Real Academia Española na versão *on-line* do Diccionario de la Lengua Española:

cachupín, na

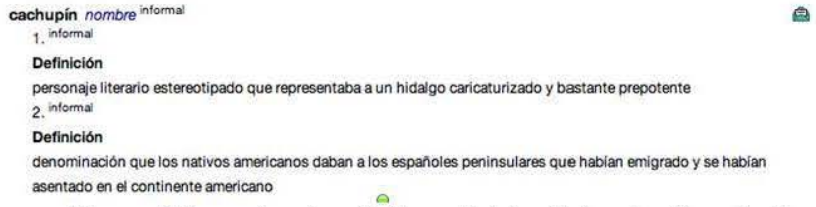
Del dim. del port. *cachopo* 'niño'.

1. m. y f. despect. **gachupín.**

Real Academia Española © Todos los derechos reservados

Figura 6: O lema "cachupín, na" no Diccionario de la Lengua Española.

Este modo de apresentar as duas palavras pode ser, com razão, considerado sexista, como se a palavra feminina *cachupina* fosse apenas subordinada à palavra masculina *cachupín*. Nos Dicionários On-line de Espanhol Valladolid-UVa, as duas palavras são tratadas separadamente e listadas como dois lemas diferentes. Isso também sugere que cada uma das duas palavras foi pesquisada separadamente no Google, que busca itens de significado. O resultado surpreendente desse método é que a palavra masculina aparece com dois sentidos, enquanto a palavra feminina inclui nada menos que oito sentidos, ou seja, um total de dez sentidos, como pode ser visto nas capturas de tela do dicionário extraído do banco de dados e apresentadas nas Figuras 7 e 8.



cachupín *nombre* informal

1. informal

Definición

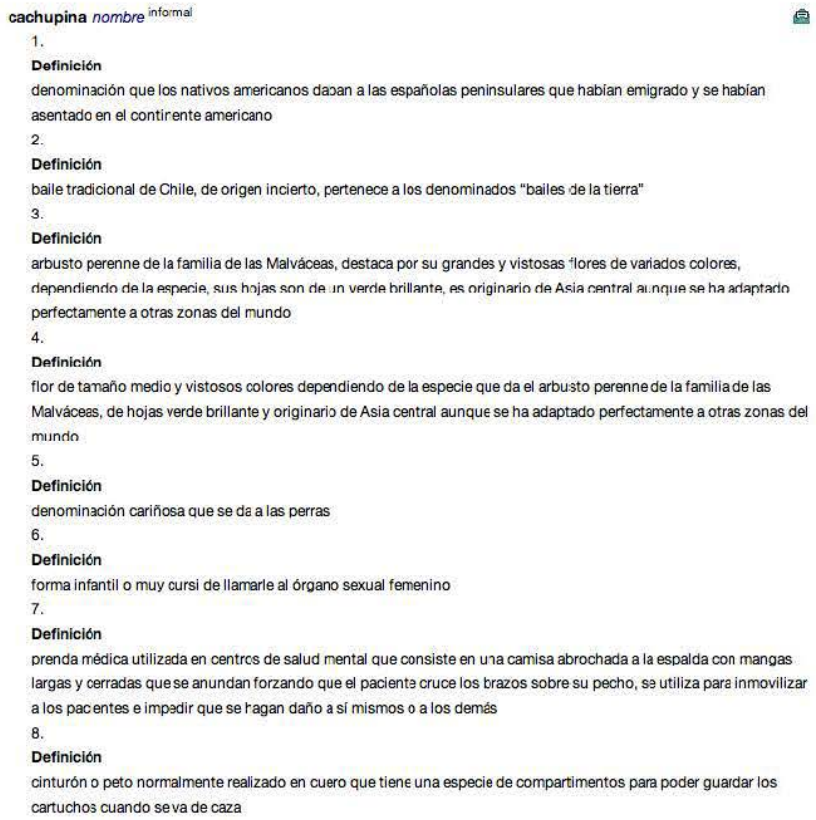
personaje literario estereotipado que representaba a un hidalgo caricaturizado y bastante prepotente

2. informal

Definición

denominación que los nativos americanos daban a los españoles peninsulares que habían emigrado y se habían asentado en el continente americano

Fig. 7: O lema "cachupín" nos Dicionários On-line de Espanhol Valladolid-UVa.



cachupina *nombre* informal

1.

Definición

denominación que los nativos americanos daban a las españolas peninsulares que habían emigrado y se habían asentado en el continente americano

2.

Definición

baile tradicional de Chile, de origen incierto, pertenece a los denominados "bailes de la tierra"

3.

Definición

arbusto perenne de la familia de las Malváceas, destaca por su grandes y vistosas flores de variados colores, dependiendo de la especie, sus hojas son de un verde brillante, es originario de Asia central aunque se ha adaptado perfectamente a otras zonas del mundo

4.

Definición

flor de tamaño medio y vistosos colores dependiendo de la especie que da el arbusto perenne de la familia de las Malváceas, de hojas verde brillante y originario de Asia central aunque se ha adaptado perfectamente a otras zonas del mundo

5.

Definición

denominación cariñosa que se da a las perras

6.

Definición

forma infantil o muy cursi de llamarle al órgano sexual femenino

7.

Definición

prenda médica utilizada en centros de salud mental que consiste en una camisa abrochada a la espalda con mangas largas y cerradas que se anudan forzando que el paciente cruce los brazos sobre su pecho, se utiliza para inmovilizar a los pacientes e impedir que se hagan daño a sí mismos o a los demás

8.

Definición

cinturón o peto normalmente realizado en cuero que tiene una especie de compartimentos para poder guardar los cartuchos cuando se va de caza

Figura 8: O lema "cachupina" nos Dicionários On-line de Espanhol Valladolid-UVa.

As diferenças acima não devem ser absolutizadas, pois o número de lemas e sentidos coletados no banco de dados dos Dicionários On-line de Espanhol Valladolid-UVa não pode ser comparado diretamente com os outros quatro dicionários do espanhol.

Há duas razões para isso. A primeira é que os quatro dicionários analisados são impressos, e, por isso, sofrem com as restrições de espaço. A segunda é que a "filosofia do papel" continua a influenciar a atual lexicografia espanhola, mesmo no ambiente virtual. Em relação a isso, um problema sério é que os critérios de seleção ainda não foram adaptados à nova tecnologia. Rundell (2015) mostra, de forma condensada, como esses critérios mudaram frente a desenvolvimentos recentes:

> Então, quando não há restrições de espaço, pode ser que faça mais sentido mudar a pergunta e, em vez de questionar, "esta palavra passa nos meus testes de inclusão?", devamos perguntar, "há razões plausíveis para não incluir essa palavra?" (Rundell 2015: 312)

Esta mudança nos critérios de seleção sugere que os lexicógrafos, que, há apenas algumas décadas, deviam justificar a inclusão de qualquer novo lema ou sentido - porque, normalmente, significava a exclusão de outros dados -, são agora desafiados com a necessidade de justificar a não-inclusão de novos lemas e acepções. A esse respeito, não sabemos quantas acepções ainda não publicadas os outros dicionários espanhóis podem ter em seus bancos de dados. Também não sabemos quantos itens de significado adicionais poderiam ser identificados em seus *corpora*, se fosse necessário. A comparação entre os cinco dicionários apresentados na Seção 6 deve ser compreendida como um indicativo das novas possibilidades que o uso da Internet coloca à disposição da lexicografia, e não como um fato baseado em evidências que explica todos os aspectos. No final do dia, o que deve ser comparado não são os dicionários em si, mas as bases e os métodos empíricos usados para fornecer os dados lexicográficos.

Com isso em mente, nossa hipótese atual, em fase de submissão para outras pesquisas, é que um *corpus* composto de textos selecionados é mais apropriado para identificar as palavras, os sentidos e os comportamentos das palavras mais comuns e mais frequentes, enquanto o uso da Internet diretamente como um *corpus* é um método mais apropriado quando há a necessidade de detectar palavras, expressões, sentidos e comportamentos menos típicos e frequentes das palavras.

7. Qualidade e produtividade

Como resultado da discussão anterior, conclui-se que o uso da Internet diretamente como um *corpus* representa um método promissor para explorar dois resultados relevantes provindos do desenvolvimento tecnológico, a saber: 1) que a Internet hoje compreende um número quase "ilimitado" de textos e palavras; 2) que um dicionário digital moderno é sustentado por um banco de dados com capacidade de armazenamento quase "ilimitada". A escolha do método é, portanto, um tópico interessante para discussões acadêmicas e uma questão que tem consequências práticas e econômicas consideráveis. Atualmente, o desafio para as editoras e equipes lexicográficas não é apenas compilar dicionários de alta qualidade, mas também garantir alta produtividade no

processo de compilação. Cada vez mais, os lexicógrafos estão se dando conta de que esse ofício está imerso em uma crise que, de certa maneira, pode ser descrita como uma luta entre a vida e a morte. Essa crise é determinada por duas tendências opostas na lexicografia atual: por um lado, muitos editores de dicionários de alta qualidade estão fechando seus departamentos de dicionário devido à falta de renda e a um modelo de negócios sustentável. Por outro lado, um número crescente de dicionários de livre acesso, mas de qualidade duvidosa, é colocado na Internet por pessoas geralmente bem intencionadas mas sem treinamento suficiente.

Os resultados dessa evolução são muitos e, em sua maioria, negativos. Embora os dicionários, devido à Internet, tenham mais usuários do que nunca, muitos desses usuários enfrentam problemas quando tentam utilizar as informações retiradas de dicionários duvidosos, um fato que muitos professores de línguas e de tradução reconhecerão. O resultado inevitável é que um número cada vez maior de usuários, conscientes de suas necessidades, recusa a lexicografia e procura outras ferramentas para obter informações. Algumas grandes empresas espanholas que estão dispostas a pagar pelo serviço são, por exemplo, críticas ao padrão atual de dicionários *on-line* de espanhol, deixando um mercado aberto para projetos como o Valladolid-UVa. No entanto, o problema é ainda mais exacerbado pelo fato de que muitos usuários da Internet, especialmente os jovens, esperam que o serviço seja gratuito, como discutido por Gouws & Tarp (2017).

Editoras de todo o mundo estão lutando para encontrar uma solução para esses complexos desafios. Uma das contra-medidas necessárias para a crise atual é, sem dúvida, aumentar a produtividade no processo de compilação do dicionário, visando reduzir custos e encontrar um modelo de negócio sustentável – ou seja, aumentar a produtividade sem comprometer a qualidade. A produtividade tem muitas faces e só pode crescer por meio da integração de tecnologia amigável, de métodos eficientes e de lexicógrafos bem treinados e motivados. Em projetos de dicionários modernos como o Valladolid-UVa, a interface do lexicógrafo é a ferramenta central de trabalho por meio da qual eles introduzem os dados no banco de dados. A esse respeito, Tarp (2015) escreve:

> A interface do lexicógrafo é basicamente um meio de produção. Por conseguinte, deve ser concebida visando garantir uma produtividade elevada e a mais alta qualidade possível do produto, ou seja, dos dados armazenados na base de dados. Para isso, é necessário, acima de tudo, que contenha todos os campos necessários para introduzir dados lexicográficos dos tipos previstos na base de dados. Mas também é importante que a interface seja tão amigável quanto for possível, a fim de facilitar o trabalho do lexicógrafo, reduzir o número de erros, economizar nos recursos empregados e encurtar o tempo total de produção. (Tarp 2015: 234)

As Figuras 2, 4 e 5 da Seção 6 mostram três capturas de tela da interface do lexicógrafo relacionada ao lema *cachupina*. O primeiro representa o cartão-mãe no qual os dados gramaticais comuns a todos os sentidos do lema são introduzidos. A segunda e terceira captura de tela mostram duas páginas do cartão representando o primeiro sentido do lema. As páginas são estruturadas de acordo com as diferentes tarefas que devem ser executadas e de tal forma que a necessidade de alternar entre uma e outra por meio dos botões funcionais à esquerda seja reduzida ao máximo. A ideia principal é que os lexicógrafos se sintam confortáveis ao trabalhar com essa interface. Seu *design* de fácil utilização, juntamente com os métodos descritos nas seções anteriores, é condição para a alta produtividade que caracteriza a compilação dos Dicionários On-line de Espanhol Valladolid-UVa.

A identificação de itens de significado e a introdução de definições e outros dados lexicográficos no banco de dados começaram em março de 2014. A experiência mostra que um lexicógrafo pode finalizar uma média de 4 a 6 acepções por hora com o método descrito e a ferramenta de produção. A experiência também mostra que a produtividade diminui após quatro ou cinco horas, pois o trabalho exige um alto grau de concentração – razão pela qual os quatro lexicógrafos que fazem essa parte do trabalho trabalham apenas quatro horas diárias no projeto. Mas, se abstrairmos esse fato, isso significaria que um lexicógrafo em tempo integral em um dia de trabalho de oito horas seria capaz de terminar cerca de 40 acepções e em uma semana de 40 horas cerca de 200 acepções, o que resulta em um total de cerca de 9.000 acepções por lexicógrafo em um ano com carga horária de 45 semanas.

O resultado é que os quatro lexicógrafos vinculados ao projeto, trabalhando por meio turno, concluíram um total de cerca de 40.000 registros (cartões) de março de 2014 a julho de 2016. Uma comparação poderia ser feita com o pequeno exército de lexicógrafos que está trabalhando no dicionário da Real Academia Española e que, acreditamos, conta com cerca de 20 pessoas. Se essa equipe trabalhasse com a ferramenta e o método descritos, seriam capazes de produzir cerca de 180.000 acepções por ano. Se trabalhassem durante três anos, passariam para mais de meio milhão de acepções e, passados apenas cinco anos e meio, o número atingiria um milhão de acepções, ou seja, o dicionário de espanhol mais completo já produzido. Em relação a isso, o desafio está feito! Basicamente, trata-se de se adaptar e explorar plenamente as novas tecnologias e técnicas colocadas à disposição da lexicografia. Os Dicionários On-line de Espanhol Valladolid-UVa fornecem um exemplo de como isso pode ser feito, embora não afirmemos que este é o único caminho que leva à Roma.

8. Competências e papel ativo do lexicógrafo

Apenas ferramentas, métodos e bases empíricas não constituem um dicionário por si só. Por mais avançada que seja a tecnologia, o mais importante na criação de dicionários ainda é o fator humano, através de lexicógrafos habilidosos, experientes e motivados. Dicionários, como Gudmann (2014: 31) afirma corretamente, "ainda são feitos por seres

humanos reais através de um processo criativo, sem que exista uma resposta absolutamente correta".

Então, o que exatamente é exigido dos lexicógrafos que participam do projeto? Aqui, mais uma vez, é necessário fazer uma distinção entre conhecimento e habilidades. Isto significa, por um lado, que um projeto *on-line* moderno, como os Dicionários On-line de Espanhol Valladolid-UVa, não pode prosperar sem um gerente (lexicógrafo-chefe) que tenha um profundo conhecimento da teoria e da metodologia lexicográfica, bem como a capacidade de projetar um conceito de dicionário, escrever instruções, selecionar e treinar uma equipe de colaboradores e supervisionar o trabalho diário.

Por outro lado, também significa que o projeto precisa de uma equipe de lexicógrafos habilidosos, altamente produtivos e capazes de gerar dados lexicográficos com a qualidade desejada. Esses profissionais devem, acima de tudo, ter competências linguísticas em espanhol, ou seja, devem ser falantes nativos da língua espanhola. Além disso, eles também devem ter "um bom conhecimento do mundo em geral e sobre, pelo menos, uma área específica" (Bergenholtz 2013: 5). O conhecimento específico pode ser sobre linguística, mas também pode ser sobre outras disciplinas relevantes para o projeto. A esse respeito, Bergenholtz (2013) relata que a equipe de lexicógrafos que trabalha nos dicionários dinamarqueses *on-line* é composta por pessoas de estudos linguísticos, da matemática, da química, da biologia molecular, da física, das ciências jurídicas, da economia e da química. Quando os lexicógrafos espanhóis foram testados antes de serem empregados no projeto Valladolid-UVa, além da competência e do conhecimento da língua, tinham o dever de mostrar outras habilidades relevantes, como a capacidade de usar computadores, navegar na Internet, encontrar dados relevantes de acordo com as instruções e transformar esses dados em definições facilmente compreensíveis na língua espanhola.

No entanto, este é apenas o ponto de partida. Em algumas das ações listadas na Seção 5, fica claro que a seleção de itens de significado não é uma ciência exata com "uma resposta absolutamente correta". Pelo contrário, um resultado bem-sucedido depende, em grande parte, do papel e das decisões ativas do lexicógrafo, que não se baseiam apenas nas competências e nos conhecimentos de linguagem, mas também na experiência. Pelo menos, este é o caso das seguintes ações:

- Ação 4: Quantas páginas devem ser ignoradas?
- Ação 6: Quais partes dos mini-textos são relevantes?
- Ação 8: Quantas páginas devem ser revisadas?
- Ação 9: Quando a quantidade de dados empíricos é suficiente?
- Ação 11: Quando o lexicógrafo está satisfeito com o processo?

As decisões tomadas nesses casos afetarão a qualidade do produto final e o próprio tempo para tomar a decisão terá consequências na produtividade. O desafio é, portanto, reduzir o tempo de tomada de decisão e elevar a qualidade das decisões. Para entender o que acontece, ou deveria acontecer, é útil referir-se aos "cinco estágios de aquisição de

habilidades" propostos por Dreyfus & Dreyfus (1986), que operam com os seguintes tipos de estágios de acordo com as habilidades: novato, iniciante avançado, competente, proficiente e experiente. Flyvbjerg (2001) resumiu as características dessas cinco etapas no processo de aprendizagem:

> (1) Os novatos agem com base em elementos e regras independentes do contexto. (2) Iniciantes avançados também usam elementos situacionais, que aprenderam a identificar e interpretar com base em sua própria experiência em situações semelhantes. (3) Os competentes são caracterizados pela escolha de metas e planos como base para suas ações. Metas e planos são usados para estruturar e armazenar informações dependentes e independentes do contexto. (4) Os proficientes identificam problemas, metas e planos intuitivamente a partir da própria perspectiva baseada na experiência. A escolha intuitiva é verificada pela avaliação analítica antes da ação. (5) Finalmente, o comportamento dos experientes é intuitivo, holístico e sincrônico, compreendido na maneira como determinada situação libera um quadro de problema, meta, plano, decisão e ação em um instante e sem divisão em fases. Este é o nível da verdadeira expertise humana: os experientes são caracterizados por um desempenho fluente e sem esforço, livre de deliberações analíticas. (Flyvbjerg 2001: 20-21)

Flyvbjerg (2001: 21) acrescenta que o modelo acima contém um "salto qualitativo" dos três primeiros estágios para os estágios 4 e 5, e que "o salto implica um abandono do pensamento baseado em regras como a base mais importante para a ação, e sua substituição por contexto e intuição". Em outra publicação, Dreyfus e Dreyfus (1992) enfatizam esse ponto:

> Parece que os iniciantes fazem julgamentos usando regras e características rígidas, mas, com talento e muita experiência, o iniciante se torna um especialista que vê intuitivamente o que fazer sem precisar aplicar regras e fazer julgamentos. A tradição intelectualista resultou em uma descrição precisa do iniciante e do especialista diante de uma situação desconhecida, mas normalmente um especialista não deliberou os resultados. Ele não raciocina. Ele nem sequer age deliberadamente. Ele simplesmente faz espontaneamente o que normalmente funcionou e, portanto, normalmente funciona. (Dreyfus e Dreyfus 1992: 117)

Se esse modelo for transferido para a lexicografia, será mais fácil entender o que faz um bom lexicógrafo e o que deve ser levado em conta ao selecionar uma equipe de colaboradores para um projeto de dicionário.

No caso concreto do projeto Valladolid-UVa, quando o anúncio de emprego foi publicado, mais de 90 candidatos demonstraram interesse. Destes, 12 candidatos foram pré-selecionados com base em seus conhecimentos e competências documentados. Foi oferecido a eles (fevereiro de 2014) um curso de 30 horas ministrado por Pedro A. Fuertes-Olivera e Helene R. Gudmann, uma lexicógrafa habilidosa com experiência em dicionários de Internet dinamarqueses. O curso incluiu a introdução à lexicografia e ao projeto Valladolid-UVa, além de instruções sobre como coletar dados na Internet, escrever definições e preparar as categorias de dados restantes. Os 12 candidatos fizeram um teste no qual deveriam preencher um número de cartões no banco de dados com base nas instruções. Os quatro melhores profissionais foram selecionados e começaram a trabalhar por um período experimental de três meses - uma exigência da legislação espanhola.

No momento do teste, os quatro lexicógrafos que, posteriormente, conseguiram o emprego poderiam ser caracterizados como *novatos* de acordo com o modelo de Dreyfus. No período experimental de três meses, esperava-se que eles se desenvolvessem relativamente rápido para *iniciantes avançados* e depois para *profissionais competentes*, uma transformação que se refletiria no aumento da produtividade e da qualidade do trabalho lexicográfico. Se não fosse o caso, o contrato seria cancelado. No entanto, após três meses, a tendência é que os profissionais sejam ainda dependentes das instruções lexicográficas (regras) e dos objetivos e planos escolhidos como base para suas ações e decisões, embora alguns deles possam começar a usar sua intuição com base em experiências anteriores. Se esse for o caso, eles saltarão para o estágio quatro, o do profissional *proficiente*, que é o nível mínimo que deveria ser esperado de todos os lexicógrafos que participam de projetos de tal magnitude e importância, como o dos Dicionários On-line de Espanhol Valladolid-UVa.

O quinto estágio no modelo de Dreyfus é o nível de *experiente*, o qual só é alcançado por uma porção de lexicógrafos praticantes. Depende, acima de tudo, da experiência, mas não pode ser alcançado sem talento, o qual deve ser identificado no período experimental. Nesse estágio, o lexicógrafo simplesmente age espontaneamente sem deliberar e, portanto, "mesmo os melhores lexicógrafos, quando pressionados, nunca conseguem explicar o que estão fazendo ou por qual razão" (Wierzbiecka 1985: 5). A lexicografia tornou-se puramente artesanal. No entanto, isso certamente não implica que "a lexicografia não tem fundamento teórico", como também afirma Wierzbiecka, ou que o processo de compilação lexicográfica não se baseia em regras. Significa, na verdade, que essas regras foram completamente internalizadas e integradas com a intuição baseada na experiência, em um desempenho fluente, holístico e que exige pouco esforço, em que o lexicógrafo, como qualquer outra pessoa que atua nesse nível, não sabe explicar, exatamente, o que está fazendo. A teoria lexicográfica e as instruções (regras) ainda se

fazem presentes, como pano de fundo, sendo uma importante e necessária instrumentária para transmitir conhecimentos e habilidades para futuros lexicógrafos.

Atualmente, os lexicógrafos que participam do projeto na Universidade de Valladolid podem ser caracterizados como executores proficientes ou experientes, conforme definido no modelo Dreyfus. Isto implica que eles agora são capazes de tomar decisões rápidas, qualificadas e intuitivas para agir em situações como as discutidas acima (Ações 4, 6, 8, 9 e 11), bem como em todas as outras situações que podem estar relacionadas ao processo de compilação. Nesse sentido, a intuição humana baseada na experiência é um importante fator de produção sem o qual o sucesso em um projeto lexicográfico seria impossível.

Essa lógica pode ser resumida da seguinte forma: em primeiro lugar, a tecnologia e os métodos necessários para trabalhar diretamente com a Internet como um *corpus* exigem seres humanos qualificados, instruídos e talentosos, motivados a ter um desempenho extraordinário; e, em segundo lugar, essas características devem ser percebidas pelo gerente de projeto o mais rapidamente possível - uma habilidade que também exige experiência e talento. Essa é, pelo menos, a experiência de trabalho atual dos Dicionários On-line de Espanhol Valladolid-UVa.

9. Conclusões

Os *corpora* foram introduzidos na década de 1960, enquanto a Internet, como fenômeno generalizado, foi disseminada até três décadas depois, na década de 1990. É surpreendente que até agora a lexicografia tenha feito mais uso de uma tecnologia antiga do que de uma mais recente. A questão é se essa seria a hora de explorar as possibilidades lexicográficas da Internet. A experiência dos Dicionários Valladolid-Uva indica claramente que o tempo é mais do que oportuno. Isso mostra que os lexicógrafos habilidosos e bem treinados que trabalham com as ferramentas e os métodos certos são perfeitamente capazes de lidar com as eventuais desvantagens ao usar a Internet diretamente como um *corpus* em vez dos tradicionais *corpora* de texto. Isso não significa, obviamente, que os *corpora* não tenham mais relevância para a lexicografia, pois ainda têm um papel importante a desempenhar relacionado a diversas tarefas. Significa, acima de tudo, que a lexicografia, para enfrentar a atual crise, precisa estar no ambiente virtual não apenas para apresentar seus produtos, mas também para deixá-los com a qualidade e a produtividade necessárias.

Neste artigo, discutimos um conjunto de dicionários *on-line* de espanhol, uma das línguas mais faladas do mundo, com mais de quatrocentos milhões de falantes nativos. É evidente que o número de textos em espanhol disponíveis na Internet é enorme. No entanto, de acordo com a experiência dos dicionários *on-line* dinamarqueses, é perfeitamente possível usar a mesma tecnologia e metodologia ao trabalhar com uma língua menos falada, com apenas cinco milhões de falantes maternos. Em relação a isso, o possível problema não é tanto o número de falantes, mas sim a penetração e o uso generalizado da Internet dentro de certa comunidade de fala. Isso sugere que pode haver

algumas línguas africanas com relativamente poucos falantes, em que a coleta de textos da Internet ainda não é grande o suficiente para compilar dicionários, conforme aqui descrito – mas elas serão a exceção à regra. Na maioria dos casos, a quantidade de textos da Internet já é suficiente, ou será no futuro próximo. A Internet está aqui para ficar, pelo menos por alguns anos, e seria um grande erro não começar a explorar hoje suas possibilidades lexicográficas[8].

Referências

Dicionários

BERGENHOLTZ, H. Ed. 2016. *De Danske Netordbøger*. Odense: Ordbogen.com.

FUERTES-OLIVERA, P.A.; H. BERGENHOLTZ (Eds.) in collaboration with M.Á. Sastre Ruano, E. Álvarez Ramos, M. Fonseca Herández, M.J. López Carrero, Á. Prieto Salvador and O. Saldaña. *Diccionarios en Línea de Español "Universidad de Valladolid"*. Hamburg: Lemma.com. (Under construction).

MOLINER, M. 2007. *Diccionario de Uso del Español*. Third Edition. Madrid: Gredos.

Real Academia Española. 2014. *Diccionario de la Lengua Española*. 23rd Edition. Madrid: Espasa.

SÁNCHEZ PÉREZ, A. Ed. 2001. *Gran Diccionario de Uso del Español Actual*. Madrid: Sociedad General Española de Librería.

SECO, M.; O. ANDRÉS; G. RAMOS. 2011. *Diccionario del Español Actual*. Madrid: Aguilar.

Leituras adicionais

ATKINS, B.T.S.; M. RUNDELL. 2008. *The Oxford Guide to Practical Lexicography*. Oxford, New York: Oxford University Press.

BERGENHOLTZ, H. 1996. Korpusbaseret Leksikografi. *LexicoNordica* 3: 1-15.

BERGENHOLTZ, H. 2013. The role of Linguists in Planning and Making Dictionaries in Modern Information Society. D. Kwary, N. Wulan and L. Musyahda. Eds. *Lexicography and Dictionaries in the Information Age. Selected papers from the 8th ASIALEX Conference*: 1-10. Surabaya: Airlangga University Press.

BERGENHOLTZ, H. ; H. AGERBO. 2014. Meaning Identification and Meaning Selection for General Language Monolingual Dictionaries. *Hermes* 52: 125-139.

[8] N.A.: Agradecemos ao Ministério da Economia e Competitividade espanhol pelo financiamento do projeto "La Teoría Funcional de la Lexicografía: Diseño y Construcción de Diccionarios de Internet" (Ref. FFI2014-52462-P).

BERGENHOLTZ, H.; S. TARP. Eds. 1995. *Manual of Specialised Lexicography*. Amsterdam, Philadelphia: John Benjamins.

DREYFUS, H.; S. DREYFUS. 1986. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Free Press.

DREYFUS, H.; S. DREYFUS. 1992. What is Moral Maturity? Towards a Phenomenology of Ethical Expertise. J. Ogilvy. Ed. *Revisioning Philosophy*: 111-131. Albany: State University of New York Press.

FLYVBJERG, B. 2001. *Making Social Science Matter. Why Social Inquiry Fails and How It Can Succeed Again*. Cambridge: Cambridge University Press.

FRANCIS, W.N. 1979. Problems of Assembling and Computerizing Large Corpora. H. Bergenholtz and B. Schaeder. Eds. *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora*: 110–123. Königstein/Ts.: Scriptor.

FUERTES-OLIVERA, P.A. 2012. Lexicography and the Internet as a (Re-)source. *Lexicographica* 28: 49-70.

FUERTES-OLIVERA, P.A.; H. BERGENHOLTZ. 2015. Los Diccionarios en Línea de Español "Universidad de Valladolid". *Estudios de Lexicografía* 4: 71-98.

FUERTES-OLIVERA, P.A.; S. TARP. 2014. *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography*. Berlin, Boston: De Gruyter.

GOUWS, R.H.; S. TARP. 2017. Information Overload and Data Overload in Lexicography. *International Journal of Lexicography*. 30 (4), 389-415.

GUDMANN, H.R. 2014. *Betydningshuller i Spanske Ordbøger. En Undersøgelse af Betydningsenheder i Spanske Monolingvale Almene Receptionsordbøger*. Master Thesis. Aarhus: Aarhus University, Department of Business Communication.

GUDMANN, H.R. 2015. Lagunas de Significado en los Diccionarios Españoles. *Estudios de Lexicografía* 4: 161-184.

HANKS, P. 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25 (4): 398-436.

KILGARRIFF, A. 1997. I Don't Believe in Word Senses. *Computers and the Humanities* 2 (31): 91–113.

KILGARRIFF, A. 2012. [Review of] Pedro A. Fuertes- Olivera/Henning Bergenholtz (Eds.). e-Lexicography: The Internet, Digital Initiatives and Lexicography. *Kernerman Dictionary News*, July 2012: 26-29.

KILGARRIFF, A.; G. GREFENSTETTE. 2003. Introduction to the Special Issue on the Web as Corpus. Computational *Linguistics* 29: 333-347.

LEES, R. 1962. Oral contribution. Quoted by Francis (1979): 110.

RUNDELL, M. 2015. From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. *Lexikos* 25: 301-322.

SINCLAIR, J.M. 1997. Introduction. *Collins Cobuild English Language Dictionary.* London: HarperCollins Publishers, xv-xxi.

TARP, S. 2008. *Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography.* Tübingen: Niemeyer.

_____. 2015. Structures in the Communication between Lexicographer and Programmer: Database and Interface. *Lexicographica* 31: 17-46.

_____. 2016. Excesos en el Uso de Corpus en la Lexicografía: «Pesca» de Términos y Definiciones. *Revista de Lexicografía* 21: 145-163.

TARP, S.; P.A. FUERTES-OLIVERA. 2016. Advantages and Disadvantages in the Use of Internet as a Corpus: The Case of the Online Dictionaries of Spanish Valladolid-Uva. *Lexikos* 26: 273-295.

WIERZBIECKA, A. 1985. *Lexicography and Conceptual Analysis.* Ann Arbor: Karoma.

XUE, M.; S. TARP. 2016. Corpus-based, Corpus-driven or Corpus-assisted lexicography? The Limited Usefulness of Corpora in Defining Specialised Terms. *Lexicographical Studies*, 4: 1-11.

Como citar este texto (ABNT):
TARP, S.; FUERTES-OLIVEIRA, P.A. Tradução de Luísa Rabaldo. Métodos e técnicas para usar a Internet diretamente como *corpus*: o caso dos Dicionários on-line de Espanhol Valladolid-UVa. **Cadernos de Tradução**, Porto Alegre, n. 43, jul/dez, p. 10-32, 2018.